Using the OntoGene pipeline for the triage task of BioCreative 2012

Rinaldi, Fabio ; Clematide, Simon ; Hafner, Simon ; Schneider, Gerold ; Grigonyte, Gintare ; Romacker, Martin ; Vachon, Therese

Abstract: In this article, we describe the architecture of the OntoGene Relation mining pipeline and its application in the triage task of BioCreative 2012. The aim of the task is to support the triage of abstracts relevant to the process of curation of the Comparative Toxicogenomics Database. We use a conventional information retrieval system (Lucene) to provide a baseline ranking, which we then combine with information provided by our relation mining system, in order to achieve an optimized ranking. Our approach additionally delivers domain entities mentioned in each input document as well as candidate relationships, both ranked according to a confidence score computed by the system. This information is presented to the user through an advanced interface aimed at supporting the process of interactive curation. Thanks, in particular, to the high-quality entity recognition, the OntoGene system achieved the best overall results in the task.

DOI: https://doi.org/10.1093/database/bas053

DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2013, Article ID bas053, doi:10.1093/database/bas053

Original article

Using the OntoGene pipeline for the triage task of BioCreative 2012

Fabio Rinaldi[1,*], Simon Clematide[1], Simon Hafner[1], Gerold Schneider[1], Gintarė Grigonytė[1], Martin Romacker[2] and Therese Vachon[2]

[1]Institute of Computational Linguistics, University of Zurich, Binzmuhlestrasse 14, Zurich 8050, Switzerland and [2]Novartis Pharma AG, NIBR-IT, Text Mining Services, Basel, Switzerland

***Corresponding author:** Tel: +41 79 300 67 71; Fax: +41 44 635 68 09; Email: fabio.rinaldi@uzh.ch

Citation details: Fabio Rinaldi, Simon Clematide, Simon Hafner, Gerold Schneider, Gintarė Grigonytė, Martin Romacker, and Therese Vachon. Using the OntoGene pipeline for the triage task of BioCreative 2012. *Database* (2012) Vol. 2012: article ID bas053; doi:10.1093/database/bas053.

In this article, we describe the architecture of the OntoGene Relation mining pipeline and its application in the triage task of BioCreative 2012. The aim of the task is to support the triage of abstracts relevant to the process of curation of the Comparative Toxicogenomics Database. We use a conventional information retrieval system (Lucene) to provide a baseline ranking, which we then combine with information provided by our relation mining system, in order to achieve an optimized ranking. Our approach additionally delivers domain entities mentioned in each input document as well as candidate relationships, both ranked according to a confidence score computed by the system. This information is presented to the user through an advanced interface aimed at supporting the process of interactive curation. Thanks, in particular, to the high-quality entity recognition, the OntoGene system achieved the best overall results in the task.

Introduction

As a way to cope with the constantly increasing generation of results in molecular biology, some organizations maintain various types of databases that aim at collecting the most significant information in a specific area. For example, UniProt/SwissProt (1) collects information on all known proteins. IntAct (2) is a database collecting protein–protein interactions. The Comparative Toxicogenomics Database (CTD) collects associations between chemicals and genes in order to support the study on the effects of environmental chemicals on health (3). Most of the information in these databases is derived from the primary literature by a process of manual annotation known as 'literature curation'. Text mining solutions are increasingly requested to support the process of curation of biomedical databases.

Several community-run evaluations have been organized in the past few years in order to assess the advancement of the field and stimulate new developments. Some of the best known are BioCreative (4), BioNLP (5) and CALBC (6). The 2012 BioCreative edition included, in particular, a task aiming at supporting the triage process for the Comparative Toxicogenomics Database. In this article, we describe the approach used for our participation in the triage task of the BioCreative 2012 challenge and the results obtained.

The triage task is the first step of the curation process for several biological databases: it aims at selecting and prioritizing the articles to be curated in the rest of the process. In BioCreative 2012, the task organizers provided a chemical entity to be used as an entry point of the curation process, and a list of articles to be prioritized according to that chemical.

Our solution to this task has been implemented under the assumption that articles should be considered relevant if they are related to the target entity provided as input and additionally, their relevance should be increased by the presence of interactions in which the target chemical is involved.

The work presented here is part of the OntoGene project (http://www.ontogene.org/), which aims at improving biomedical text mining through the usage of advanced natural

Database, Vol. 2013, Article ID bas053, doi:10.1093/database/bas053

language processing techniques. Our approach is based on accurate processing of the input articles by a pipeline of advanced NLP tools, which perform increasingly complex task, from sentence splitting and tokenization up to term recognition, phrase chunking and syntactic analysis (7, 8).

In the context of the SASEBio project (Semi-Automated Semantic Enrichment of the Biomedical Literature), the OntoGene group has also developed a user-friendly interface (ODIN: OntoGene Document INspector) which presents the results of the text mining pipeline in an intuitive fashion and allows a deeper interaction of the curator with the underlying text mining system (9).

In the rest of this article, we first explain how our existing OntoGene relation mining system has been customized for the CTD dataset ('Information extraction' section), and then how it has been integrated with a conventional information retrieval (IR) system (Lucene) for the purpose of the triage task ('Integration with a standard IR system' section). We also provide a brief overview of our ODIN curation interface ('The ODIN interface' section), an evaluation of the results obtained by the integrated sytem in the shared task ('Evaluation' section) and a discussion on current and future work ('Discussion' section).

Information extraction

In this section, we describe the OntoGene Text Mining pipeline which is used to (i) provide all basic pre-processing (e.g. tokenization) of the target documents, (ii) identify all mentions of domain entities and normalize them to database identifiers and (iii) extract candidate interactions. We then describe in detail, a machine-learning approach used to obtain an optimized scoring of candidate interactions based upon global information from the set of interactions existing in the CTD database (excluding data from the test set).

Pre-processing and detection of domain entities

The OntoGene Text Mining pipeline was used in order to transform the input documents into a richly annotated XML format, which is the basis of our relation extraction algorithm. The assumption was that from this format we could derive information useful to improve document ranking and therefore provide a solution for the triage task, which could improve on a conventional IR approach. In a previous work (10), we showed that the inclusion of PubMed metadata, such as the list of chemical substances as well as the annotated MeSH descriptors and qualifiers, improves the detection of important relations and enhances term recognition coverage. Therefore, we added such metadata from the PubMed XML files as a textual list at the end of each abstract. In the OntoGene text mining pipeline, sentence and token boundaries of the enriched abstracts are identified using the LingPipe framework (more information can be found at http://alias-i.com/lingpipe).

In this section, we describe in particular our approach to named entity recognition, i.e. the problem of detecting names of relevant domain entities in biomedical literature (genes, chemicals and diseases for CTD) and grounding them to widely accepted identifiers assigned by the original database.

Terms, i.e. preferred names and synonyms, are automatically extracted from the original CTD database and stored in a common internal format, together with their unique identifiers, as obtained from the original resource. An efficient lookup procedure is used to annotate any mention of a term in the documents with the IDs to which it corresponds. A term normalization step is used to take into account a number of possible surface variations of the terms. The same normalization is applied to the list of known terms at the beginning of the annotation process, when it is read into memory, and to the candidate terms in the input text, so that a matching between variants of the same term becomes possible despite the differences in the surface strings. In case the normalized strings match exactly, the input sequence is annotated with the IDs of the reference terms and no further disambiguation on concepts is done at this point. For more technical details of the OntoGene term recognizer, see (11).

Detection of interactions

As a baseline approach, it is possible to generate candidate interactions among domain entities on the basis of their co-occurrence in a given text span (typically one or more sentences or an even larger observation window). Such an approach might achieve a sufficient recall but suffers from low precision. In order to obtain better precision it is possible to take into account the syntactic structure of the sentence, or the global distribution of interactions in the original database. In this section, we describe in detail how candidate interactions are ranked by our system, according to their relevance for CTD curation, by exploiting the vast amount of curated articles in the CTD database.

For the entities in the CTD database a context window of one sentence for candidate relation generation is too restrictive. In an evaluation limited to those PubMed articles from CTD with explicit evidence for at most 12 relations we found the following distribution: for about 32% of all relations from the CTD, where our term recognizer was able to detect both participating entities, there was no sentence containing both entities in the PubMed abstract. Given these numbers, we chose to use a context window of the entire abstract for candidate pair generation.

An initial ranking of the candidate relations can be generated on the basis of frequency of occurrence of the respective entities only:

$$\text{relscore}(e_1, e_2) = [f(e_1) + f(e_2)]/f(E)$$

where $f(e_1)$ and $f(e_2)$ are the number of times the entities e_1 and e_2 are observed in the abstract, while $f(E)$ is the total count of all identifiers in the abstract. Previous experiments for the extraction of protein–protein interactions from PubMed abstracts (8) and more recent experiments on the PharmGKB database (12) have shown that giving a 'boost' of ~10 to the entities contained in the title produces a measurable improvement of ranking of the results.

This simple approach can be further optimized if we apply a supervised machine-learning method for scoring the probability of an entity to be part of a relation which was manually curated and inserted into the CTD database. There are two key motivations for this approach. First, we need to lower the scores of false positive relations which are generated by too broad entities (frequent but not very interesting). The goal is to model some global properties of the curated CTD relations. Second, we want to penalize false positive concepts that our term recognizer detects. In order to deal with such cases, we need to condition the entities by their normalized textual form t. The combination of a term t and one of its valid entities e is noted as $t : e$.

For example, according to the term database of the CTD, the word 'PTEN' (phosphatase and tensin homolog) may denote nine different diseases (autistic disorder; carcinoma, squamous cell; glioma; hamartoma syndrome, multiple; head and neck neoplasms; melanoma; prostatic neoplasms; endometrial neoplasms; craniofacial abnormalities), apart from denoting the gene 'PTEN'. Using the techniques described below we can automatically derive the relevancy of the concepts related to the word 'PTEN' from the corpus of manually curated CTD relations. Doing so leads to a result which clearly prefers the interpretation of 'PTEN' as a gene.

Next, we define a predicate $\text{gold}(A, e)$ which is true for an article A if there is at least one relation in the gold standard where entity e is part of and false (i.e. 0) otherwise. We estimate the overall probability $P[\text{gold}(A, e) = 1 | t : e]$ with the help of the maximum entropy modeling tool *megam* (13). For training, we use the set of CTD-referenced PubMed articles having not more than 12 manually curated relations (the threshold of 12 relations is motivated by the observation that the more relations an article has, the less probable it is to find them by processing the abstracts only), additionally removing all articles which are part of the BioCreative training and test set for the respective dataset (this results in 22319 articles for the training set, containing 69320 curated relations. For the test set, we used 22 825 articles with 71 064 relations).

For unseen normalized terms t, i.e. terms not present in the training data, the maximum entropy classifier would assign a low default probability based on the distribution of all training instances. However, we can specify better back-off probabilities if we take into account the admissible entity/entities e of term t. Our current back-off model works as follows: if the entity e of an unseen term t is seen in the article, the averaged probability of all seen term–entity pairs is used. Otherwise, the averaged probability of all entities of the same type as e is used.

The score of an entity e in an article A is the sum of all zone-boosted term frequencies (as mentioned earlier, occurrences in the title are counted 10 times) weighted by their gold probability:

$$\text{score}(e) = \sum_{t:e \in A} f(t : e) \times P[\text{gold}(A, e) = 1 | t : e]$$

Having determined the individual score for each entity e, we compute the relation score as the harmonic mean of its component scores:

$$\text{relscore}(e_1, e_2) = 2 \times \frac{\text{score}(e_1) \times \text{score}(e_2)}{\text{score}(e_1) + \text{score}(e_2)}$$

In our previous work on relation ranking (10), the relation score was taken as a sum of the concept scores. By performing systematic cross-validation experiments on all CTD articles, we noticed that using the harmonic mean improves the results considerably. In order to make the relation scores comparable between different articles we normalize all relation scores for a given BioCreative dataset. For the normalization step, all relation candidate scores of a dataset are linearly scaled to a value between 0 and 1.

Integration with a standard IR system

A conventional IR system (Lucene) is used to provide a baseline document ranking from which a classification can be derived by selection of a threshold. Information derived from the OntoGene pipeline, and from the ranking process described in the previous section, is then added as additional features in order to improve the baseline ranking generated by the IR system [the integration of the various components is performed using mainly JRuby (http://jruby.org/), through which the Lucene API is accessed].

Database, Vol. 2013, Article ID bas053, doi:10.1093/database/bas053

Terminology-aware tokenization

The IR system processes the documents in the standard way, selecting different boost values for title and abstract: 10 for title, 3 for abstract, just as in the CTD reference system (notice that the boosting mentioned here is internal to the IR system, while in the previous section we mentioned a similar boosting factor for the OntoGene pipeline). Experiments with different boost values for title and abstract did not show any statistically significant change in the MAP scores, probably because most of the information is in the abstract, not in the title: the existence of relevant information in the title typically implies relevant information in the abstract.

The only significant technical change to Lucene preprocessing is the replacement of the 'StandardAnalyzer' component (which is the default analyzer for English, responsible for tokenization, stemming, etc.) with our own tokenization results, as delivered by the OntoGene pipeline. The advantage of this approach is that we can flexibly treat recognized technical terms as individual tokens and map together their synonyms (14). In other words, after this step all known synonyms of a term will be treated as identical by the IR system.

The 'StandardAnalyzer' component is replaced by a simple transformation of the XML output of the pipeline into a format suitable for internal processing by Lucene. In particular, tokens and terms as recognized by the pipeline are transformed into Lucene 'token' data objects. Whenever a domain entity (denoted by the Term element in the XML representation) is found, it is replaced by a 'normalized' version of the token sequence (term normalization involves concatenation of the lowercase version of all tokens into a single token, plus some minor ad-hoc changes that depend on the type of the term). At the same position, a new token with the text of the concept identifier is added to the input stream as seen by the IR system.

For example:

```
<W C="VBN" id="W151" o1="758" o2="767">inhibited</W>
<Term allvalues="MESH_D015232:chem" id="TW152W153"
     matched="prostaglandine2" type="chem">
  <W C="NN" id="W152" o1="768" o2="781">prostaglandin</W>
  <W C="NN" id="W153" o1="782" o2="784">E2</W>
</Term>
<W C="NN" id="W154" o1="785" o2="794">synthesis</W>
```

will be converted to the following (square brackets denote token boundaries):

```
[inhibited] [prostaglandin E2] [synthesis]
            [MESH_D015232]
```

Synonymous terms (as identified by the pipeline) are mapped to their unique identifiers (for this experiment the term identifier provided by the CTD database), which in the example above is a MeSH term. The initial search is conducted by mapping the target chemical to the corresponding identifier, which is then used as a query term for the IR system application.

Relation-based query expansion

Participants in the shared task were not only required to provide an optimized ranking of target documents, but also to deliver other relevant entities (genes, diseases and chemicals) mentioned in each abstract. The quality of the delivered entities was used as part of the overall evaluation. As described in section 2.2, the OntoGene pipeline is not only capable of delivering an optimized tokenization, it can also be used to annotate all relevant entities and to generate candidate interactions, which can be directly used for curation purposes by CTD curators.

Although the definition of the task did not require the participants to deliver candidate interactions, we worked under the assumption that documents which contain relevant interactions would be relevant themselves. When another term is often seen in relation with the target term, it is probably important for the target. This statistical information can be used to adjust the ranking of the documents.

The organizers provided for each target chemical a set of articles to be ranked by the participants. The OntoGene pipeline delivers candidate interactions as part of its standard output for each single document. Each interaction is assigned a score in the interval (0,1].

All relations that involve a term equivalent to the target (the target or one of its synonyms) were considered. From these relations, we extracted the interacting entity (the second term in those interactions). An expanded query was then created, combining the original search term with all other entities which are seen to interact with it in the target abstract. The additional query terms are weighted according to the normalized score of the interactions from which they are extracted.

As an example, suppose two documents (Document 1 and Document 2) contain the interactions schematically represented in the first two columns below (an interaction is represented as a triple of two arguments and a probability):

Document 1	Document 2	Expansion terms with score
A C 1	**A B 1**	**C 1** *from doc 1*
B C 0.7	B D 0.42	**B 0.75** *from doc 1 (score 0.5) and doc 2 (score 1)*
A B 0.5		**D 0.4** *from doc 1*
A D 0.4		

If the target term is A, the relations marked in boldface are relevant, which gives us new search terms to be added

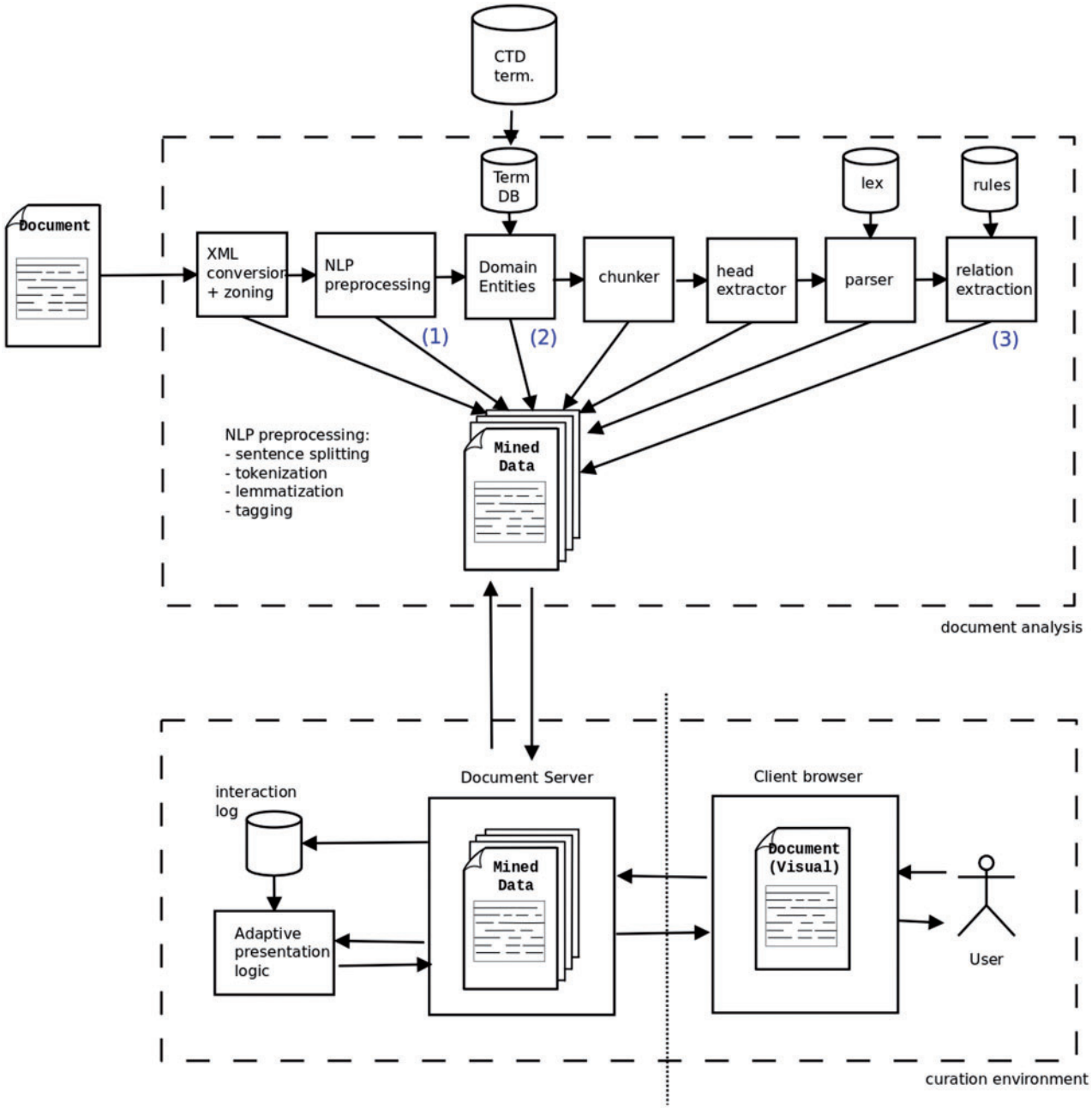


Figure 1. General architecture of the OntoGene system. The OntoGene pipeline delivers a richly annotated version of the original document. For the experiments described in this article, we made use of (i) tokens, (ii) domain entities and (iii) relations.

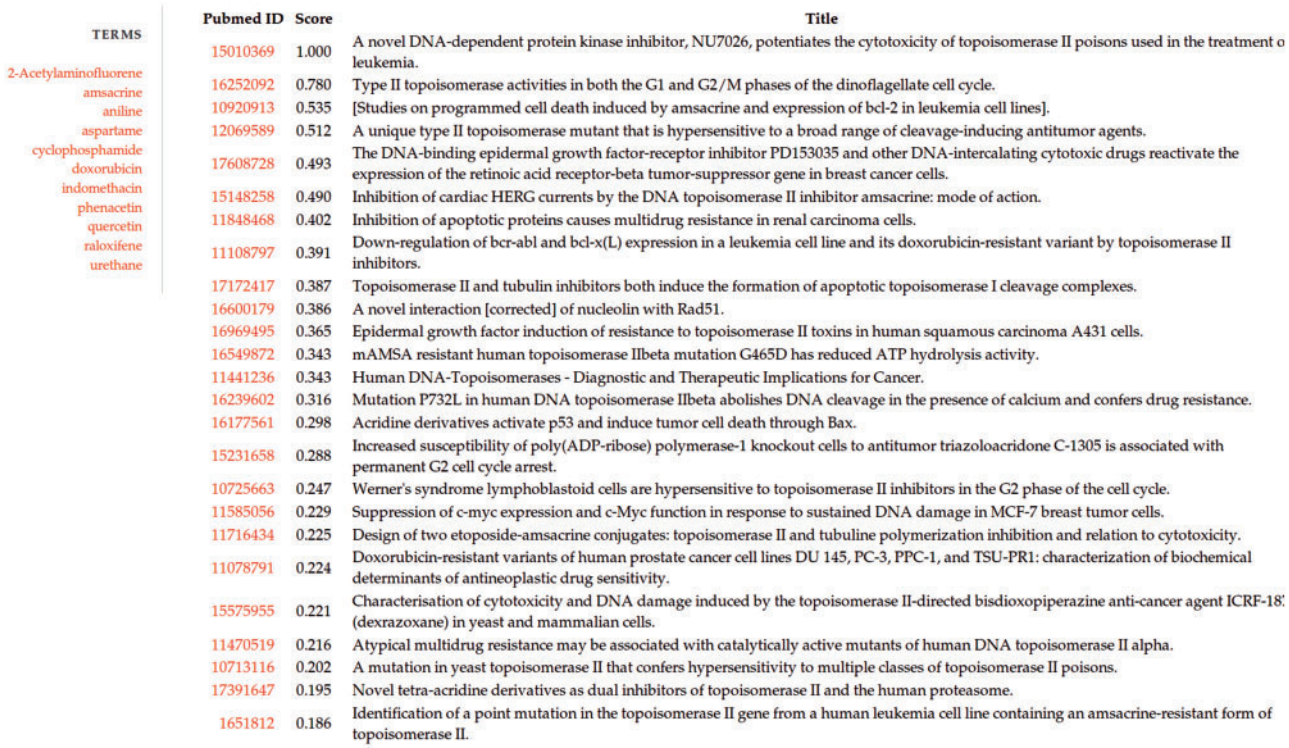
to the query, listed in the third column with their normalized weights (sum of scores divided by the number of relations).

In the search process, Lucene compares the expanded query with all the entities that are found in any given document. We have experimentally verified on the training data that this query expansion process improves the average MAP scores from 0.622 to 0.694.

The ODIN interface

The results of the OntoGene text mining system are made accessible through a curation system called ODIN, which allows a user to dynamically inspect the results of their text mining pipeline. A previous version of ODIN was used for participation in the 'interactive curation' task of the BioCreative III competition (15). This was an informal

Database, Vol. 2013, Article ID bas053, doi:10.1093/database/bas053



TERMS

2-Acetylaminofluorene
amsacrine
aniline
aspartame
cyclophosphamide
doxorubicin
indomethacin
phenacetin
quercetin
raloxifene
urethane

Pubmed ID	Score	Title
15010369	1.000	A novel DNA-dependent protein kinase inhibitor, NU7026, potentiates the cytotoxicity of topoisomerase II poisons used in the treatment o leukemia.
16252092	0.780	Type II topoisomerase activities in both the G1 and G2/M phases of the dinoflagellate cell cycle.
10920913	0.535	[Studies on programmed cell death induced by amsacrine and expression of bcl-2 in leukemia cell lines].
12069589	0.512	A unique type II topoisomerase mutant that is hypersensitive to a broad range of cleavage-inducing antitumor agents.
17608728	0.493	The DNA-binding epidermal growth factor-receptor inhibitor PD153035 and other DNA-intercalating cytotoxic drugs reactivate the expression of the retinoic acid receptor-beta tumor-suppressor gene in breast cancer cells.
15148258	0.490	Inhibition of cardiac HERG currents by the DNA topoisomerase II inhibitor amsacrine: mode of action.
11848468	0.402	Inhibition of apoptotic proteins causes multidrug resistance in renal carcinoma cells.
11108797	0.391	Down-regulation of bcr-abl and bcl-x(L) expression in a leukemia cell line and its doxorubicin-resistant variant by topoisomerase II inhibitors.
17172417	0.387	Topoisomerase II and tubulin inhibitors both induce the formation of apoptotic topoisomerase I cleavage complexes.
16600179	0.386	A novel interaction [corrected] of nucleolin with Rad51.
16969495	0.365	Epidermal growth factor induction of resistance to topoisomerase II toxins in human squamous carcinoma A431 cells.
16549872	0.343	mAMSA resistant human topoisomerase IIbeta mutation G465D has reduced ATP hydrolysis activity.
11441236	0.343	Human DNA-Topoisomerases - Diagnostic and Therapeutic Implications for Cancer.
16239602	0.316	Mutation P732L in human DNA topoisomerase IIbeta abolishes DNA cleavage in the presence of calcium and confers drug resistance.
16177561	0.298	Acridine derivatives activate p53 and induce tumor cell death through Bax.
15231658	0.288	Increased susceptibility of poly(ADP-ribose) polymerase-1 knockout cells to antitumor triazoloacridone C-1305 is associated with permanent G2 cell cycle arrest.
10725663	0.247	Werner's syndrome lymphoblastoid cells are hypersensitive to topoisomerase II inhibitors in the G2 phase of the cell cycle.
11585056	0.229	Suppression of c-myc expression and c-Myc function in response to sustained DNA damage in MCF-7 breast tumor cells.
11716434	0.225	Design of two etoposide-amsacrine conjugates: topoisomerase II and tubuline polymerization inhibition and relation to cytotoxicity.
11078791	0.224	Doxorubicin-resistant variants of human prostate cancer cell lines DU 145, PC-3, PPC-1, and TSU-PR1: characterization of biochemical determinants of antineoplastic drug sensitivity.
15575955	0.221	Characterisation of cytotoxicity and DNA damage induced by the topoisomerase II-directed bisdioxopiperazine anti-cancer agent ICRF-18 (dexrazoxane) in yeast and mammalian cells.
11470519	0.216	Atypical multidrug resistance may be associated with catalytically active mutants of human DNA topoisomerase II alpha.
10713116	0.202	A mutation in yeast topoisomerase II that confers hypersensitivity to multiple classes of topoisomerase II poisons.
17391647	0.195	Novel tetra-acridine derivatives as dual inhibitors of topoisomerase II and the human proteasome.
1651812	0.186	Identification of a point mutation in the topoisomerase II gene from a human leukemia cell line containing an amsacrine-resistant form of topoisomerase II.

Figure 2. ODIN interface: entry page.

task without a quantitative evaluation of the participating systems. However, the curators who used the system commented positively on its usability for a practical curation tasks. An experiment in interactive curation has been performed in collaboration with curators of the PharmGKB database (16, 17). The results of this experiment are described in (12), which also provides further details on the architecture of the system.

More recently, we adapted ODIN to the aims of CTD curation, allowing the inspection of PubMed abstracts annotated with CTD entities and showing the interactions extracted by our system. Once an input term has been selected, the system will generate a ranking for all the articles that might be relevant for the target term. Figure 2 shows the results provided by the system for the input chemical 'amsacrine'. The PubMed identifier and the title of each article are provided, together with the relevancy score as computed by the system. The PubMed identifier field is also an active link, which when clicked brings the user to the ODIN interface for the selected article. Figure 3 shows a screenshot of this interface.

At first access the user will be prompted for a 'curator identifier', which can be any string. Once inside, ODIN's two panels are visible: on the left the article panel, on the right the results panel. The panel on the right has two tabs: concepts and interactions. In the 'concept' tabs a list of terms/concepts is presented. Selecting any of them will highlight the terms in the article. In the 'interactions' panel the candidate interactions detected by the system are shown. Selecting any of them will highlight the evidence in the document.

All items are active. Selecting any concept or interaction in the results panel will highlight the supporting evidence in the article panel. Selecting any term in the article panel prompts the opening of a new panel on the right (annotation panel), where the specific values for the term can be modified (or removed) if needed. It is also possible to add new terms by selecting any token or sequence of tokens in the article.

Evaluation

In order to generally assess the upper limit of our relation recognition system, we evaluated the coverage of the term recognizer on all CTD-referenced articles containing at most 12 curated relations.

Table 1 describes the coverage of term recognition for concepts and relations in experimental data, and shows that we find about three-fourth of all entities. However, the upper limits for relation detection are not the same for all relation types. Relations involving chemicals have substantially lower coverage rates which seems a bit unfortunate for the CTD triage task.

Table 2 shows the final results obtained on the training (top) and test (bottom) document sets using the online

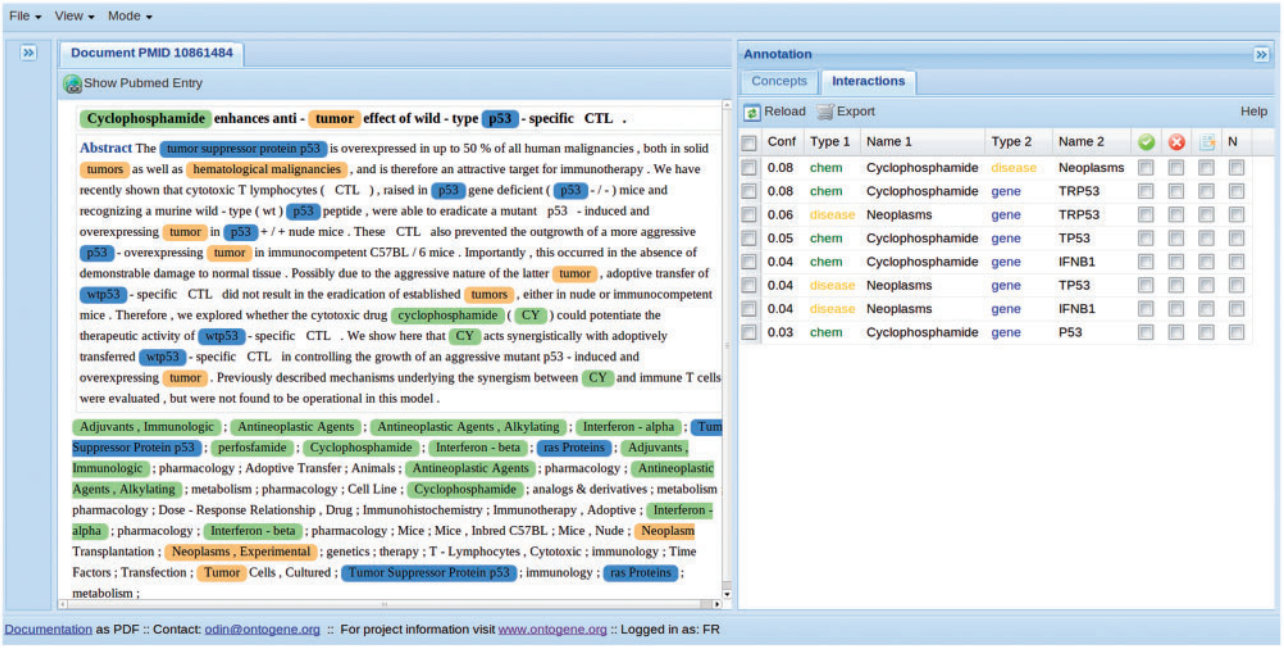


Figure 3. ODIN interface: entity annotations and candidate interactions on a sample PubMed abstract.

Table 1.

Category	Total	Found (%)
Disease	12 639	9502 (75.18)
Chemical	38 523	30 129 (78.21)
Gene	39 150	29 199 (74.58)
Total	90 312	68 830 (76.21)
dis-gen	6956	5126 (73.69)
che-dis	12 154	8356 (68.75)
che-gen	52 746	34 883 (66.13)
Total	71 856	48 365 (67.13)

Table 2.

Term	MAP	Genes	Chemicals	Diseases
Doxorubicin	0.800	0.167	0.843	0.793
Indomethacin	0.936	0.331	0.834	0.725
Raloxifene	0.798	0.244	0.818	0.778
Amsacrine	0.655	0.603	0.689	0.500
Aniline	0.543	0.625	0.561	0.524
2-Acetylaminofluorene	0.643	0.412	0.845	0.421
Aspartame	0.365	0.686	0.756	0.720
Quercetin	0.853	0.463	0.646	0.653
Cyclophosphamide	0.708	0.396	0.880	0.646
Phenacetin	0.809	0.716	0.467	0.667
Urethane	0.650	0.365	0.871	0.633

evaluation tool provided by the organizers of the shared task.

In the BioCreative 2012 shared task 1, the OntoGene pipeline proved once again its flexibility and efficiency by delivering very effective entity recognition. In particular, our system had the best recognition rate for genes and diseases and the second best for chemicals, leading to the overall best results, as can be seen in Figure 4 (18) [reproduced with permission from the author]. The query expansion approach used in combination with a standard IR system in order to generate the final article ranking did not perform as well in the test phase as the result of the training phase would have suggested. This might have been caused by overfitting to the training data.

Discussion

The OntoGene text mining pipeline provides an efficient system for the extraction of entities and relationships from the biomedical literature, as shown by the results discussed in the previous section. Additionally, the ODIN curation interface provides an user-friendly environment for the integration of information derived from the text mining tools into a curation framework.

The OntoGene system has not only been successful in several community-organized evaluations, but it has also been applied in an industrial context, within the NIBR-IT unit of Novartis Pharma AG. At Novartis, scientific

Database, Vol. 2013, Article ID bas053, doi:10.1093/database/bas053

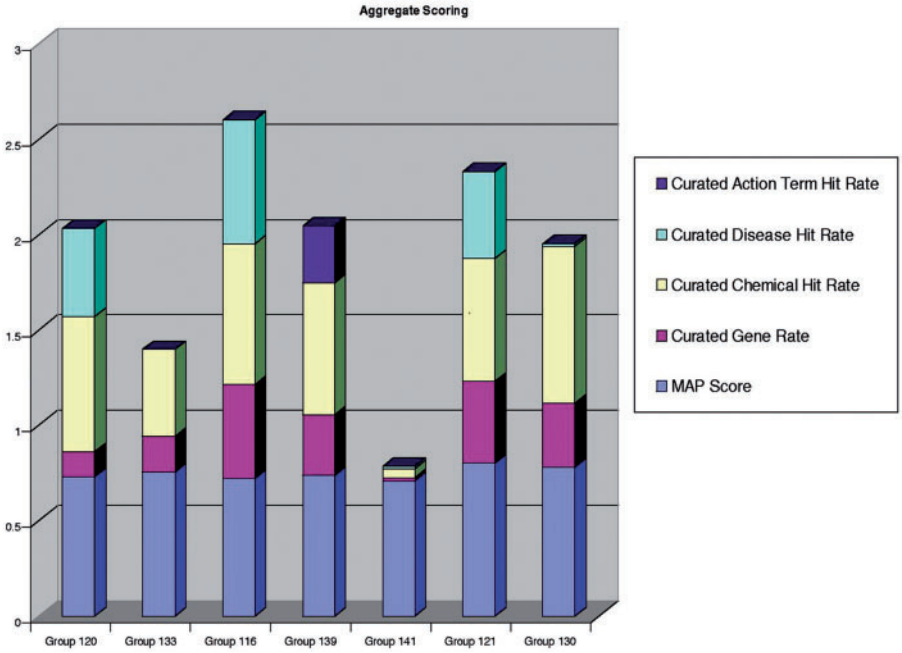


Figure 4. Official results of the BioCreative 2012 competition (task 1: 'triage for the CTD database'). OntoGene was identified as 'Group 116'. Reproduced from (18).

annotation is gaining more and more importance. In most recent applications the usage of controlled vocabularies has become mandatory. However, most of the data are still in legacy systems. This is the reason why curation of legacy data and documentation is of crucial importance. Currently, a major focus is being placed on Metadata recovery and the curation of a large variety of data repositories containing valuable knowledge in terms of assay data, scientific documentation or clinical data. The main business driver behind this initiative is that the company has a treasury of knowledge but cannot make use of it because the data are not semantically normalized.

The NIBR-IT unit of Novartis has been using ODIN to annotate textual data from legacy repositories. This application could highly benefit from the fact that the Ontogene framework is open and can easily be customized. This allows the usage of internal terminologies for lexical extraction. The legacy documents were pre-annotated with a customized pipeline and the results displayed using ODIN. The ODIN graphical user interface allows for the verification and falsification of annotation results by selecting or deselecting identified concepts. In addition, new terms can be added manually to the annotations, they can be assigned to the appropriate concept class and then fed into controlled vocabularies thus improving the extraction results of the next annotation cycle.

One of the limitations of the text mining system described above is that it does not provide the type of the detected interactions. This can be a shortcoming for some applications. For example, in the BioCreative 2012 triage task, the capacity of the system to provide a 'curated action term' was one of the factors contributing to the overall result.

The OntoGene system performs a complete syntactic analysis of each sentence in the input documents. In most cases, it is relatively easy to recover from such analysis the information which is necessary to provide a relation type. For example, Figure 5 shows a simplified representation of the analysis of the sentence 'The neuronal nicotinic acetylcholine receptor alpha7 (*nAChR alpha7*) may be involved in cognitive deficits in Schizophrenia and Alzheimer's disease.' from PubMed abstract 15695160. This sentence expresses two interactions between a gene (*nAChR*) and the diseases Schizophrenia and Alzheimer. From the graphical representation, it can be intuitively seen that the word which indicates the interaction verb 'involved' can be recovered as the uppermost node at the intersection of the syntactic paths leading to the arguments. Interaction verbs can then be used to infer a suitable CTD action code.

Table 3 shows the highest scored head words from a small subset of 93 CTD documents. The table legend explains how the various factors which contribute to the final score (rightmost column) are computed. Notice that the value 'P' is often >1, as it is not a probability value, but a relative score.

The head words in Table 3 have a high correspondence to the trigger words used in annotation tasks which use relation labels, such as BioNLP [3]. They contain few false

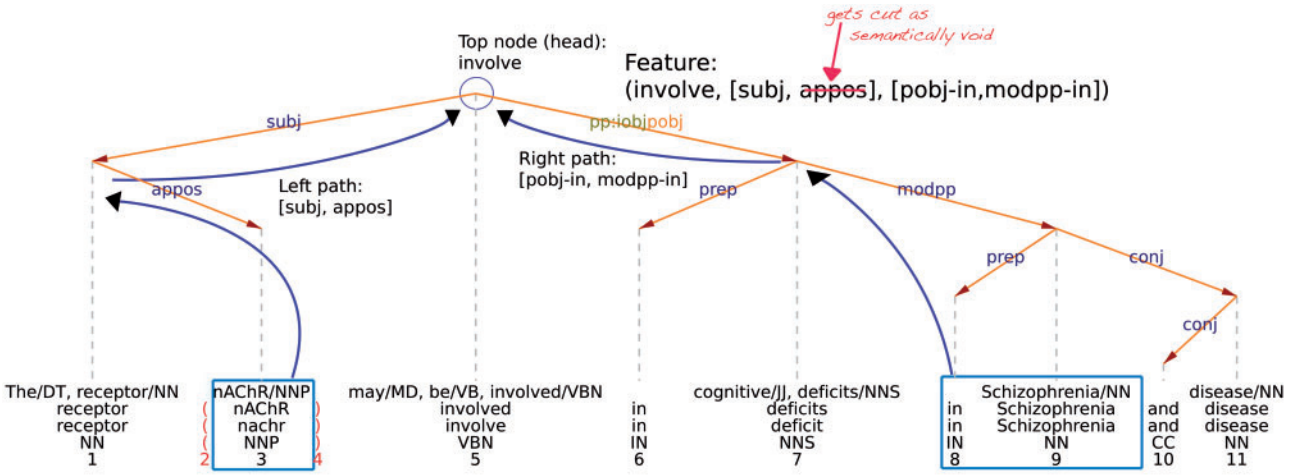


Figure 5. Example of syntactic analysis of a sentence as performed by the Ontogene parser. Reprinted from *Journal of Biomedical Informatics*, Volume 45, Issue 5, Fabio Rinaldi, Gerold Schneider, Simon Clematide, 'Relation Mining Experiments in the Pharmacogenomics Domain', pages 851–861, 2012, with permission from Elsevier.

Table 3.

Head	Term	*F* = *f*(Head)	*A* = *f*(All)	*P* = *F*/*A*	log(*F*) · log(*A*) · P.¬term
Play	0	25	17	1.47	13.41
Treat	0	24	17	1.41	12.71
Bind	0	18	9	2.00	12.70
Inhibit	0	41	48	0.85	12.28
Constitute	0	13	3	4.33	12.21
Demonstrate	0	30	30	1.00	11.57
Exhibit	0	16	11	1.45	9.67
Reveal	0	20	19	1.05	9.29
2t	0	11	4	2.75	9.14
...	...	...	...	...	...
Quinine	1	8	1	8.00	0.00
Phytoestrogen	1	7	6	1.17	0.00
Thalidomide	1	6	15	0.40	0.00

Relation labels are shown in the first column. The second column is a boolean value indicating whether the head word is itself a term. The third column ('F') shows the number of times the head word is seen in a relevant path (notice that the same head word can occur in multiple relevant paths). The fourth column ('A') shows the number of times the word occurs in the document collection. The next column shows the ratio among the preceding two values. The final column calculated a weighted score considering the previous factors.

positives (e.g. '2t' in Table 3), and they can often be mapped well to CTD action codes. For example, 'bind', 'inhibit', 'reduce', 'block', 'downregulate', 'metabolize', 'expression', 'activate', 'regulate', 'express' map to CTD action codes or BioNLP labels. Many heads refer to the investigator's conclusion ('demonstrate', 'show', 'assess', 'find', 'reveal', 'explain', 'suggest') or to methodology ('treat', 'exhibit'). Some are underspecified (e.g. 'play' which comes from 'play a role in'), and some are only syntactic operators (e.g. 'appear', 'ability'). Some are semantically ambiguous: for example, 'contribute' can equally be part of an investigator's conclusion or a syntactic operator (e.g. 'contributes to the activation'). The process of mapping these values into CTD action codes will require biological expertise for completion.

Conclusions

In this article, we have described our approach towards ranking biomedical abstracts for the triage task of the CTD curation process. The characteristic of the approach is that it gives priority to the identification of candidate interactions, which are then used as additional weighting factors in a conventional IR-based system.

The OntoGene pipeline is capable of delivering all information relevant to CTD curation: entities with their database references, interactions, and interaction terms. In the shared task, however due to insufficient time for customization, we decided to exclude the computation of interaction terms. The results of the system are accessible through an intuitive interactive interface, which will be further customized for CTD curation.

Acknowledgements

We wish to thank the anonymous reviewers for their valuable suggestions.

Funding

The Swiss National Science Foundation (grant 100014-118396/1); Novartis Pharma AG, NIBR-IT, Text Mining Services, Switzerland.

Conflict of interest. None declared.

References

1. UniProt Consortium. (2007) The universal protein resource (uniprot). *Nucleic Acids Res.*, **35**, D193–D197.
2. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C. *et al*. (2004) IntAct: An open source molecular interaction database. *Nucleic Acids Res.*, **32 (**Suppl. 1), D452–D455.
3. Mattingly,C.J., Rosenstein,M.C., Colby,G.T. *et al*. (2006) The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *J. Exp. Zool. A Comp. Exp. Biol.*, **305**, 689–692.
4. Krallinger,M., Vazquez,M., Leitner,F. *et al*. (2011) The protein-protein interaction tasks of biocreative iii: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics*, **12 (**Suppl. 8), S3.
5. Cohen,B.K., Demner-Fushman,D., Ananiadou,S. *et al*. (eds). (2009) *Proceedings of the BioNLP June 2009 Workshop*. Association for Computational Linguistics, Boulder, Colorado.
6. Rebholz-Schuhmann,D., Yepes,A., Li,C. *et al*. (2011) Assessment of ner solutions against the first and second calbc silver standard corpus. *J. Biomed. Semantics*, **2 (**Suppl. 5), S11.
7. Rinaldi,F., Schneider,G., Kaljurand,K. *et al*. (2006) An environment for relation mining over richly annotated corpora: The case of GENIA. *BMC Bioinformatics*, **7 (**Suppl. 3), S3.
8. Rinaldi,F., Kappeler,T., Kaljurand,K. *et al*. (2008) OntoGene in BioCreative II. *Genome Biol.*, **9 (**Suppl. 2), S13.
9. Rinaldi,F., Clematide,S., Garten,Y. *et al*. (2012) Using ODIN for a PharmGKB re-validation experiment. *Database*, 2012: article ID bas021; doi:10.1093/database/bas021.
10. Clematide,S. and Rinaldi,F. (2012) Ranking relations between diseases, drugs and genes for a curation task. *J. Biomed. Semantics*, **3 (**Suppl. 3), S5.
11. Rinaldi,F., Kaljurand,K. and Saetre,R. (2011) Terminological resources for text mining over biomedical scientific literature. *J. Artif. Intel. Med.*, **52**, 107–114.
12. Rinaldi,F., Schneider,G. and Clematide,S. (2012) Relation mining experiments in the pharmacogenomics domain. *J. Biomed. Inform.*, **45**, 851–861.
13. Hal Daumé,III. *Notes on CG and LM-BFGS optimization of logistic regression.* http://www.umiacs.umd.edu/~hal/docs/daume04cg-bfgs.pdf and http://hal3.name/megam/. (5 December 2012, date last accessed).
14. Rinaldi,F., Dowdall,J., Hess,M. *et al*. (2002) Terminology as knowledge in answer extraction. In: *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE02),* Nancy, France, 28–30 August 2002, pp. 107–113.
15. Arighi,C., Roberts,P., Agarwal,S. *et al*. (2011) Biocreative iii interactive task: an overview. *BMC Bioinformatics*, **12 (**Suppl. 8), S4.
16. Klein,K.E., Chang,J.T., Cho,M.K. *et al*. (2001) Integrating genotype and phenotype information: An overview of the PharmGKB project. *Pharmacogenomics J.*, **1**, 167–170.
17. Sangkuhl,K., Berlin,D.S., Altman,R.B. and Klein,T.E. (2008) PharmGKB: Understanding the effects of individual genetic variants. *Drug Metabol. Rev.*, **40**, 539–551.
18. Wiegers,T.C., Davis,A.P. and Mattingly,C.J. (2012) Collaborative biocuration-text mining development task for document prioritization for curation. *Database*, article ID bas037; doi:10.1093/database/bas037.